Identification de scripteur par une loi puissance

Rudolf Pareti, Nicole Vincent

Identification de scripteur par une loi puissance

Rudolf Pareti – Nicole Vincent

Laboratoire Crip5 - Université Paris Descartes
45 rue des saints pères 75006 Paris

rudolf@pareti.org, nicole.vincent@math-info.univ-paris5.fr

Résumé *L'identification de scripteur est un problème difficile, il peut être considéré à plusieurs niveaux en fonction des applications, on peut s'attacher à l'analyse de détails caractéristiques ou plus généralement à l'aspect de l'écriture. Nous allons présenter une nouvelle méthode pour indexer et identifier des documents anciens tels que des lettres ou des manuscrits. Notre méthode se base sur une modélisation de la répartition des fréquences des motifs présents dans l'écriture. La relation habituellement utilisée dans des domaines mono-dimentionnels est étudiée ici en analyse d'images. Nous utiliserons les paramètres caractérisant la loi mise en évidence pour indexer les manuscrits. Nous prouverons la pertinance de ces paramètres pour la reconnaissance de scripteur.*

Mots-clés *: :* identification de scripteur, méthode globale, loi puissance.

1 Introduction

Un texte est constitué d'un ensemble de lettres arrangées de façon non-aléatoire, cet ordre, constituant des mots, permet de donner du sens au texte. La communication entre les hommes est passée très tôt par l'écrit un des moyens les plus usités pour communiquer à travers le temps et l'espace. Mais au-dela du sens, des informations non sémantiques sont contenues dans l'aspect même que peut prendre l'écriture, en particulier pour les manuscrits. En effet chaque personne à son écriture propre, plus ou moins penchée, plus ou moins parallèle au bord de la feuille. Cette écriture est le reflet de sa personnalité, une personnalité qui va au-delà de celle aquise avec la culture et l'éducation. Il est tout à fait possible de reconnaître d'un seul coup d'œil l'écriture d'un proche même si celui-ci écrivait dans une langue inconnue. On pourrait penser que la reconnaissance de scripteur n'est qu'un problème moderne, il est certe important de nos jours de pouvoir identifier ou authentifier l'auteur d'un écrit ou d'une signature. Les techniques de numérisation et les logiciels de retouche d'image de plus en plus répandus rendent la contrefaçon de plus en plus facile et l'authentification de scripteur peut-être un moyen de palier ce problème. Mais la sécurité n'est pas le seul domaine à pouvoir bénéficier des avancées de la reconnaissance de scripteur. Depuis les débuts de l'Histoire, l'écriture manuscrite est utilisée pour communiquer entre les hommes, un laissé passer écrit de la main d'un roi ou une missive rédigée par un gouvernant conférait au porteur des privilèges importants. Une correspondance entre deux personnages historiques peut nous éclairer sur tel ou tel événement. D'autres problèmes peuvent apparaître avec les documents anciens, les informations peuvent avoir été détruites par le passage du temps et il devient impossible de connaître l'auteur d'une lettre, certains documents peuvent aussi ne jamais avoir renfermé une telle information, par exemple une lettre à un ami, un message sur un bout de papier écrit au coin d'une table, un bout de pièce de théatre rédigé au moment où l'inspiration fut présente. Le style global d'une écriture peut donner plus d'information à propos de qui est l'auteur d'un document que le sens même du texte, il permet d'affirmer si un document est un original ou une copie réalisée par un secrétaire.

Les expérimentations de notre méthode sont réalisées à partir d'une base de manuscrits dont nous connaissons l'identité des scripteurs, ces manuscrits ont été par la suite divisés en deux groupes. Le premier nous servira de base d'apprentissage et le second sera celui de test. Notre objectif est de montrer qu'une analyse assez globale du texte et l'extraction de primitives associées au scripteur, fournit une information assez discriminante pour l'identification des scripteurs. Les manuscrits étudiés ici sont du 16^{è} siècle.

2 Etat de l'art

De nombreuses études traitent de l'identification de scripteur, nous allons présenter ici rapidement les principaux groupes de méthodes que nous classons en trois groupes, les approches contextuelles, les approches non contextuelles et celles plus récentes qui utilisent le style l'écriture pour en reconnaître l'auteur.

2.1 Les approches contextuelles

Ces approches se basent non seulement sur l'image d'un texte mais aussi sur le contenu [1-2]. Elles se révèlent dans notre cas peu adaptées du fait qu'elles nécessitent que l'on dispose d'un texte particulier écrit par les scripteur que l'on voudrait reconnaître. Les textes étudiés ici sont de la main de personnes depuis longtemps disparues. De plus les documents sont d'une grande hétérogénéité ce qui nuit à l'éfficacité de ces méthodes.

2.2 Les approches non-contextuelles

Ces approches imposent moins de contraintes au scripteur que celles précédemment citées, elles s'appuient par exemple sur l'étude d'histogramme [3], ou une description analytique de l'écriture [4], ou encore sur un réseaux de neurones [5]. Elles implémentent des méthodes locales et ne retiennent pas le fait que les humains n'ont, dans la plupart des cas, pas besoin d'analyser longuement et en détail un texte pour en reconnaître son auteur.

2.3 Identification par le style

Le but de ces méthodes est d'extraire d'un texte des caractéristiques indépendantes de la sémantique. Un ensemble de paramètres est utilisé [6]. Les études ont souvent eu pour objectif d'adapter une méthode de reconnaissance au scripteur même. Par exemple les modèles HMM peuvent être utilisés dans le processus de reconnaissance. D'autres travaux extraient de l'écriture des formes invariantes qui sont utilisées pour adapter un processus de reconnaissance de l'écriture manuscrite et les formes extraites contribuent à l'authentification du scripteur [7-8]. Des recherches basées sur des approches fractales ont été faites pour classer ou identifier des auteurs de manuscrits. C'est dans cette lignée de méthodes que s'inscrivent nos propres recherches.

3 La métodologie

Avant de présenter la méthode proprement dite que nous avons utilisée pour aborder le problème de l'identification de scripteur, nous allons rappeler l'énoncé de la loi de Zipf qui a inspiré notre approche. Nous préciserons aussi comment transposer aux images une technique dédiée aux signaux monodimensionnels et comment nous avons adapté à l'étude des manuscrits les différentes conclusions.

3.1 La loi de Zipf

La loi de Zipf [10] est une loi empirique qui a été énoncée il y a une cinquantaine d'années. Elle repose sur une loi puissance. Cette loi affirme que dans un ensemble de symboles organisés de façon topologique, la distribution des fréquences des n-uplets, nommés motifs, n'est pas aléatoire. Notons les fréquences d'apparition de ces motifs M_1, M_2 … M_n, N_1, N_2 … N_n. Quand ils sont triés par ordre décroissant de la fréquence d'apparition, ils sont liés aux rangs de ces fréquences. Cette relation peut être formulée ainsi :

$$N_{\sigma(i)} = k \times i^{a} \qquad (1)$$

$N_{\sigma(i)}$ représente le nombre d'apparitions d'un motif de rang i, k et a sont des constantes. Cette loi est caractérisée par l'exposant a, k lui est plus lié à la longueur de la séquence de symboles étudiée. La complexité de la loi tient au fait qu'elle n'est pas linéaire. Néanmoins l'application d'un opérateur logarithmique conduit à une relation linéaire caractérisée elle aussi par deux paramètres.. De plus la valeur de l'exposant a peut être estimée expérimentalement par le coefficient directeur de la droite de régression approximant le graphe ($\log_{10}(i)$, $\log_{10}(N_{\sigma(i)})$) où *i=1à n*. Par la suite nous appellerons ce graphe la courbe de Zipf. Une façon de finaliser l'approximation est d'utiliser la méthode des moindres carrés. La non régularité des points dans le graphe oblige à réaliser un ré échantillonnage.

Nous allons prendre l'exemple en dimension 1 d'un signal de longueur quelconque, le signal est construit sur un alphabet comportant 2 éléments {a , b}. Pour analyser le signal on peut utiliser par exemple des fénêtres glissantes de longueur 2. Dans une telle fenêtre le nombre de motifs distincts est 4 {(aa) , (ab) , (ba) , (bb)}. La loi de Zipf gère la relation, pour le phénomène étudié entre le nombre de chacun des motifs et son rang 1, 2, 3 ou 4. Le premier tableau de la figure 1 présente plusieurs occurrences du signal où la position de la fenêtre glissante est à chaque fois indiquée. Le second tableau résume le nombre d'occurrences de chaque motif et le graphe illustre la forme de la relation existante.

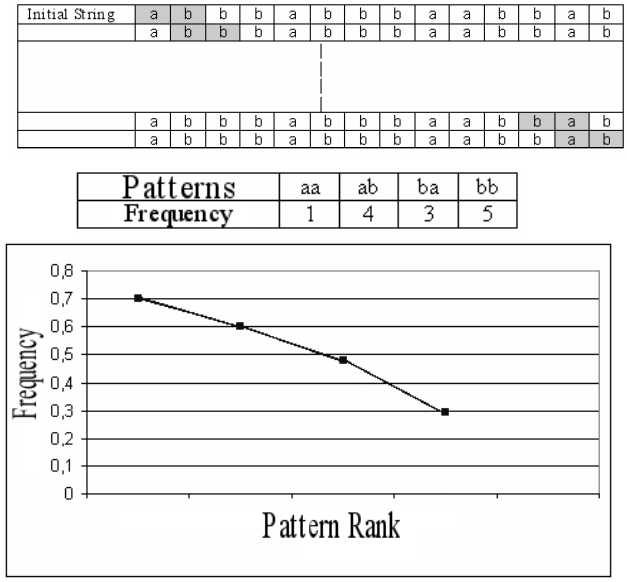


Fig. 1 – Construction de la courbe de Zipf

La validité de cette loi a été observée dans de nombreux domaines, les études portent sur des signaux mono dimensionnels [11]. Avant d'étudier l'efficacité de cette loi sur les manuscrits nous allons montrer comment adapter la loi de Zipf à l'image.

3.2 Application aux images

Les textes manuscrits que nous étudions ont été numérisés à l'aide d'un scanner en niveaux de gris où chaque pixel est encodé sur 8 bits (256 niveaux différents). L'information codée est l'intensité. Par nature un texte apparaît plutôt comme binaire (noir et blanc), nous avons préféré garder ce niveau d'information pour effectuer nos expérimentations. La qualité du papier et les déformations introduites par le coté ancien des documents nécessitent cette précision. De plus, les niveaux de gris même relatifs, ont plus de sens que les valeurs binaires.

Dans cette section nous allons montrer les différences qui apparaissent quand on traite un signal 2D, en l'occurrence des images. En effet dans le cas de signaux mono dimensionnels, les motifs observés se limitaient à

une simple succession de symboles. En ce qui concerne les images le masque du motif choisi doit respecter la topologie 2D du plan dans lequel se trouvent organisées les données. Le choix le plus naturel en ce qui concerne le masque est alors d'utiliser une matrice de 3 sur 3 qui représente un pixel et ses voisins, c'est-à-dire un voisinage du pixel.

Le principe de l'étude des fréquences des motifs reste inchangé, la relation entre le nombre d'apparitions de chaque motif et leur rang est recherchée. Néanmoins comme 256 symboles sont utilisés pour coder un pixel il y a théoriquement 256^9 motifs possibles. Ce nombre n'est pas raisonnable car souvent supérieur au nombre de pixels des images. Cela entraînerait que chaque motif serait rare et de ce fait leur fréquence aurait peu de sens d'un point de vue statistique, les fréquences perdent leur signification. Par exemple dans une image de 640x480 pixels on ne trouve que 304964 motifs. Il devient donc vital de réduire drastiquement le nombre de motifs possibles pour donner un sens au modèle. Le codage des motifs devient de ce fait la pierre angulaire du processus.

Plusieurs contraintes doivent entrer en considération pour étiqueter les motifs avec un nombre de symboles raisonnable. Quelles sont les propriétés que nous voulons mettre en avant ? Combien de classes de motif doit on considérer ? Cela ne sera résolu qu'au travers d'un nouveau codage de l'image.

3.2.1 Codage des motifs

Des études ont démontrées que la loi de Zipf était vérifiée dans le cas des images avec différents processus de codage [12]. Les images utilisées étaient des paysages ou des scènes de la vie quotidienne. Ici nous étudions des images non-naturelles que sont les manuscrits, concept élaboré par les Hommes. Nous cherchons un processus apte à discriminer ce type d'image. Notre postulat est donc que deux écritures ayant un style ressemblant doivent avoir des distributions de leurs motifs similaires. Nous avons choisi de diminuer le nombre de niveaux de gris utilisés dans l'image pour réduire le nombre de motifs possibles.

3.2.2 Quantification des niveaux de gris

Les intensités des pixels sont codées sur k niveaux de gris. Un choix judicieux de la valeur k suffit à observer une image. En poussant un peu plus en avant le raisonnement nous savons que les documents sont essentiellement de nature binaire et qu'une quantification en k classes égales conduirait automatiquement à des résultats instables dus au fait que certaines classes ne contiendraient aucun pixel. Nous avons donc décidé d'utiliser un algorithme plus adaptatif de quantification, l'algorithme de classification des k-means [13].

Niveaux de gris	0-20	21-76	77-96	97-120	
Centre	15	56	86	107	
Classe	0	1	2	3	
Niveaux de gris	121-146	147-174	175-203	204-229	230-255
Centre	133	159	190	217	242
Classe	4	5	6	7	8

TAB. 1 – Exemple de classification des niveaux de gris par la méthode des k-means

Beaucoup de méthodes s'appuient sur cet algorithme. Nous avons expérimenté plusieurs valeurs de k, le tableau 1 présente un exemple de classes obtenues en utilisant 9 classes. Bien sûr l'exemple dépend essentiellement de l'image et l'agencement des classes se base sur le nombre de pixels de chaque niveau de gris. Les classes sont de tailles variées. Nous avons vérifié expérimentalement que c'est avec 3 classes que nous obtenions les meilleurs résultats. Ceci s'explique facilement par la nature même des images étudiées qui sont fondamentalement binaires, on retrouve donc les classes associées au fond et à l'écrittrue ainsi qu'une classe rassemblant les pixels que l'on pourrait qualifier d'ambigus.

3.3 Construction de la courbe de Zipf

En fonction du codage et du contenu de l'image, l'allure générale de la courbe de Zipf peut varier énormément. Visuellement, quand la courbe de Zipf n'est pas une droite, on peut affirmer que la loi puissance n'est pas vérifiée, la loi de Zipf n'est pas vérifiée. Néanmoins, la courbe peut être approximée par un ensemble de segments de droite. Alors, plusieurs phénomènes sont mis en évidence et leur structure est quantifiée par la pente du segment correspondant.

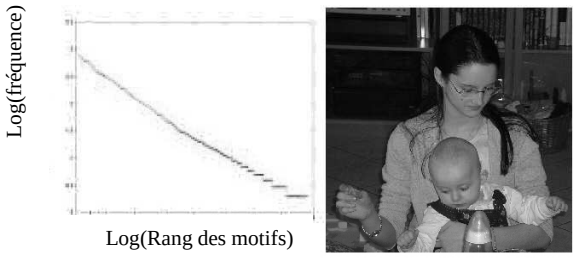


FIG. 2 – Courbe de Zipf associée à une image

Quelque soit le codage utilisé, on peut construire la courbe de Zipf. Un exemple de telle courbe est illustré par la figure 2.

3.4 Application aux images de manuscrits.

Nous allons maintenant appliquer cette construction aux manuscrits. Les observations montrent que la loi de Zipf n'est pas respectée pour les manuscrits. Mais les

courbes peuvent être approximées par quelques segments de droite. Ces zones peuvent être interprétées. Ainsi certains segments donnent des informations sur les régions tandis que d'autres nous donnent des informations sur les contours présents dans l'image. En fonction de l'écriture présente dans l'image, ces parties prennent plus ou moins d'importance et caractérisent l'image. Nous pouvons ainsi extraire des indications de structures à des niveaux différents au sein de l'image.

Nous avons pris en compte trois zones linéaires dans chaque courbe. Elles sont extraites automatiquement par un processus récursif. Un premier piont de coupure est obtenu comme le point le plus éloigné au sens de la distance euclidienne usuelle de la droite joignant les points de fréquence maximum et minimum. Ensuite le processus est itéré sur la portion droite de la courbe. Un exemple de résultat représentatif est sur la figure 3.

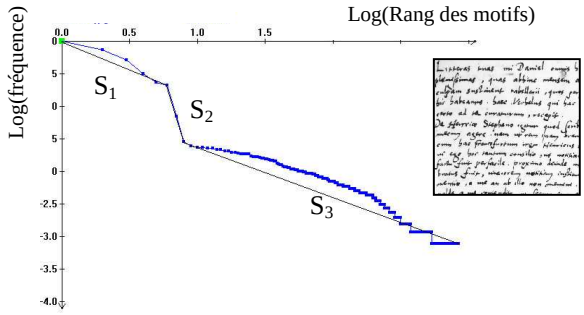


FIG. 3 – Courbe de Zipf d'un manuscrit avec extraction de zones linéaires

Ainsi sont extraites trois valeurs associées à chaque image qui correspondent aux pentes des trois zones extraites. Chacune des zones traduit une complexité différente dans la répartition des motifs qui interviennent dans le trait et des familles de niveaux de détails, plutot de contours quand on atteint des motifs peu fréquents.

4 Similarité entre les textes

Nous allons définir dans cette section les mesures effectuées et présenter les résultats des expérimentations que nous avons effectuées.

4.1 Comparaison

Nous avons choisi d'indexer les images de manuscrits avec les trois valeurs extraites des trois lois puissances mises en évidence. Ces pentes se révèleront insuffisantes pour caractériser à elles seules un scripteur, en effet la même structure peut prendre plus ou moins d'importance dans l'écriture de chacun, une propriété est plus ou moins accentuée. Il était important de pouvoir quantifier cette information dans chaque écriture. C'est pourquoi nous avons décidé d'introduire dans notre étude la notion de la longueur des segments approximant la courbe de Zipf. Nous tenons compte des abscisses des points de rupture de l'approximation caractérisant la courbe de Zipf d'une image. Un scripteur est donc représenté par 3 ou 6 valeurs.

Pour définir la distance entre deux images nous utiliserons la distance de Hamming quelque soit l'espace dans lequel on considère la représentation.

$$\text{distance}(I, I') = \sum_{i=1}^{d} \left| S_i - S'_i \right| \qquad (2)$$

4.2 Evaluation

Notre étude nous conduit à réaliser deux applications potentielles.

- A partir d'un manuscrit l'utilisateur peut demander à ce que soient affichés les manuscrits contenus dans la base qui sont les n plus semblables à la requête.
- A partir d'un manuscrit inconnu, dont on ne connaît pas l'identité du scripteur, l'application peut indiquer l'auteur correspondant, si un texte écrit par sa main est déjà contenu dans la base. La décision repose alors sur l'algorithme des k plus proches voisins.

4.2.1 Recherche et extraction d'image

La figure 4 est une illustration de notre première application avec la présentation des cinq images les plus semblables. Le test présenté a été réalisé avec l'algorithme des k-means k étant égal à 3. Les textes retrouvés ne sont pas tous du même scripteur les derniers sont d'auteurs différents de celui de l'image requête mais ils ont un style très proche difficile à différencier même à l'œil nu.

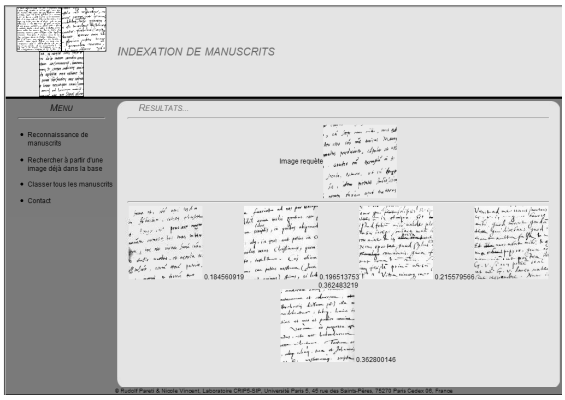


FIG. 4 – 5 images les plus similaires à l'image requête

4.2.2 Reconnaissance de scripteur

Une évaluation dépend évidemment de la nature et du nombre des documents contenus dans la base de données. Nos documents sont tous du 16ème siècle et les styles d'écriture sont variés.

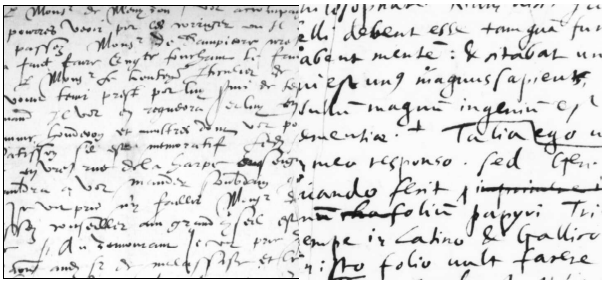
On obtient un taux de reconnaissance de 62% avec l'utilisation du motif 3x3, l'utilisation d'une représentation des textes dans l'espace de dimension 3 associé aux pentes et de l'algorithme des k-means avec k=3 et un KPPV.

K=	1	3
Taux	55%	62%

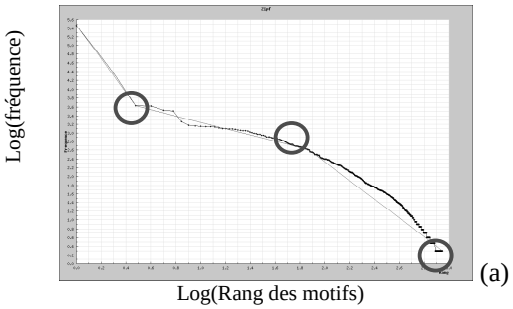
TAB. 2 – Taux de reconnaissance de scripteur par la méthode des k ppv

En analysant les erreurs on s'aperçoit que les manuscrits mal reconnus sont proches en style et en regardant de plus près leur courbe de Zipf on se rend compte que les pentes sont approximativement les

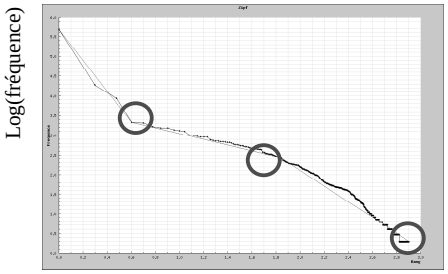
mêmes mais que la longueur des segments varie. C'est pourquoi pour améliorer les résultats qui sont dus à un manque de discrimination nous avons introduit les abscisses de rupture de pentes comme indiqué en figure 5.



(a) (b)



(a)



(b)

FIG. 5 – Exemple d'images proches mais d'auteurs différents ainsi que les courbes de Zipf associées.

En ajoutant ces nouvelles caractéristiques dans la comparaison avec la distance de Hamming les résultats deviennent bien meilleurs comme on peut le constater dans le tableau 3.

K=	1	3
Rate	68%	80%

. TAB. 3 – Taux de reconnaissance de scripteurs par la méthode des k ppv

5 Loi de zipf inverse

Une seconde loi a été mise en évidence par Zipf toujours dans le domaine des textes, c'est la loi de Zipf inverse. Il a déjà été observé qu'elle s'appliquait dans le domaine des images. Cette loi prend la mêm forme que la loi de Zipf directe, c'est à dire qu'elle met en jeu une loi puissance caractéisée par son exposant. Par contre, ce sont les fréquences et les nombres de motifs ayant cette fréquences qui sont mis en relation selon la formule

$$N(f) = h \times f^{b} \tag{3}$$

Cette loi n'est en fait vérifiée que pour les faibles valeurs de la fréquence. Dans notre étude nous nous sommes limités aux 10 fréquences les plus faibles. La valeur de l'exposant, toujours calculée comme le coefficient directeur de la droite de regression approximant l'ensemble des points expérimentaux, peut être utilisé comme index caractéristique d'une écriture et d'un scripteur. La figure 6 montre un exemple de courbe de Zipf inverse sur laquelle on voit bien la qualité de la linéarité de la courbe obtenue.

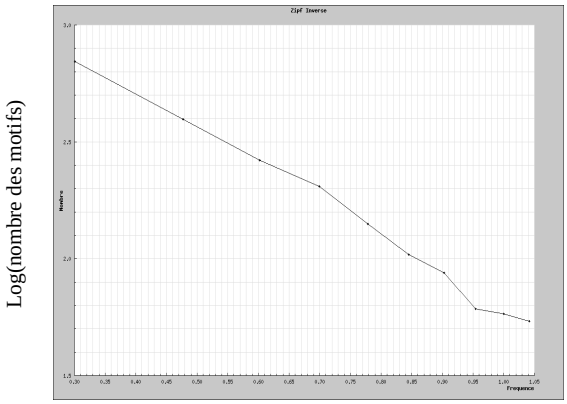


FIG. 6 – Exemple de courbe de Zipf inverse associée à l'image du document de la figure 5 (a).

Ce nouveau paramètre utilisé seul ne peut évidemment donner de résultats convenables, en revanche les motifs concernés par les deux lois ne sont pas les mêmes. Les indices extraits fournissent donc des informations complémentaires.

Les résultats avec cette nouvelle méthode sont meilleurs car les documents d'un même scripteur apparaissent dans les 2 premiers voisins les plus proches.

6 Conclusion

Certes les résultats que nous présentons ici ne sont pas optimums mais ils sont plus qu'encourageants. De plus les documents que nous avons posent des contraintes fortes de part la nature de leur digitalisation, ils viennent de photos imprimées puis scannées. Le fait que ces documents aient été reliés conduit aussi à des dégradations fortes car une partie de la photo est courbée. De plus le temps et ses méfaits ont aussi œuvrés et certains documents sont très dégradés.

On peut de ce fait affirmer que le modèle développé pour les signaux mono dimensionnels peut être adapté aux images et en particulier aux manuscrits. Les méthodes de codage des motifs mises en évidence ici ne sont pas exhaustives et les expérimentations de nouveaux processus peuvent être expérimentés. On peut aussi penser que les approches explicitées ici ne se bornent pas aux manuscrits et pourraient être appliqués à d'autres types ou parties de documents anciens ou non.

On peut noter aussi l'invariance des résultats à la rotation ce qui est essentiel pour ce type de document.

Références

[1] F. Mihelic, N. Pavesic, L. Gyergyek, "Recognition of writers of handwritten texts", *International Conference On Crime Countermeasures*, p 237-240 1977

[2] R.-D. Naske, "Writer recognition by prototype related deformation of handprinted characters", *ICPR New –York 1982* p 819-822

[3] B. Arazi, "Handwriting identification by means of runlength measurements", *IEEE Transactions on Systems, Man and Cybernetics*, SMC-7, n°12, p878-881 Dec. 1977

[4] W. Kuckuck, B. Rieger, K. Steinke, "Automatic writer recognition", *Proceedings of the 1979 Carnahan Conference on Crime Countermeasures, Lexington, Kentucky, USA* p 57-64 1979

[5] U.-V. Marti, R. Messerli, H. Bunke, « Writer identification using text line based features » *ICDAR 2001 USA* p 101-105

[6] M. Gilloux, "Writer adaptation for handwritten word recognition using hidden Markov models", *ICPR 1994 USA* p 135-139 vol. 2

[7] A. Nozary, L. Heutte, T. Paquet, Y. Lecourtier, "Difining writer's invariants to adapt the recognition task", *ICDAR '99, Bangalore, India*, pp. 765-768. 1999.

[8] A. Bensefia, A. Nozary, L. Heutte, T. Paquet, "Writer identification by writer's invariants" *IWFHR'02, Niagara on the Lake, Canada*, pp. 274-279, 2002.

[9] N. Vincent, V. Bouletreau, R. Sabourin,H. Emptoz, "How to use fractal dimensions to qualify writings and writers", *Revue Fractals, World Scientific*, Vol 8, n°1, p 85-97, 2000.

[10] G.K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, 1949

[11] E Dellandréa, P. Makris, N. Vincent, "Wavelets and Zipf Law for Audio Signal Analysis", *7th International Symposium on Signal Processing and its Applications (ISSPA 2003), Paris (France)*, Vol. 2, p. 483-486, Juil. 2003.

[12] Y. Caron, H. Charpentier, P. Makris, N. Vincent, Power Law Dependencies to Detect Regions Of Interest, *11th International Conference DGCI 2003, Naples, Italy, November 2003.*

[13] J. A. Hartigan, M. A. Wang, "K-mean clustering Algo", *JSTOR revue* p 100-108.